Research Paper

Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

Annotation Based Fast Navigation of Web-Data Retrieval

Engineering

KEYWORDS : Data Annotation, Web Databases, data alignment, data filtering, frequency annotation, multimode text.

Amit Kumar Yadav | Asst. Prof. Department of Computer Science & Application, SAC, Jabalpur

ABSTRACT

Annotation of web pages is an area of research which is getting lot of attention as the count of websites of specific topics and as a whole is increasing very fast. Since all the databases are accessible over web through HTML representations and data extraction over web is becoming more and more dynamic. Such data is huge and for applications such as online shopping comparison, article collection etc. Annotation of such collected information leads to several advantages including fast decision making, relevant information visiting, to reduce the time of futile searches, historical data management and elimination of older searches.

This paper is intended to provide an insight of the annotation techniques and application of few techniques to provide the required results with the above stated advantages. Works of various researchers in the field of annotating data has been more on limited tokens and focus is on creating dynamic annotations only. This work proposes to apply dynamic annotations on web sites data with tokenization done using all sort of tokens including long text having no specific tokens.

For machine learning and training frequency based annotations, common knowledge annotators and schema value annotators are being applied which are going to facilitate for correct annotation process. For annotation website pages shall be looked for content type, presentation style, data type, tag path and adjacencies of the contents.

1. INTRODUCTION

An annotation is metadata, a comment, and explanation, presentational markup, attached to text, image, or other data. There are many ways annotations are used by the persons for different purposes. Annotations refer to a specific part of the original data.

Such as students often highlight or underline passages in books in order to refer back to key phases easily, or add marginalia to aid studying. One educational technique when analyzing the prose literature is to have students or teachers circle the names of characters and put rectangular boxes around phrases identifying the setting of a given scene.

Similarly annotated bibliographies add commentary on the relevance or quality of each source, in addition to the usual bibliographic information that merely identifies the source.

Various media has different ways for annotations either available inherently or they can be involved as per the requirements of the users. E.g. Markup languages like XML and HTML annotate text in a way that is syntactically distinguishable from that text. They can be used to add information about the desired visual presentation, or machine-readable semantic information.

The "annotate" function (also known as "blame" or "praise") used in source control systems such as Team Foundation Server and Subversion determines who committed changes to the source code into the repository. This outputs a copy of the source code where each line is annotated with the name of the last contributor to edit that line (and possibly a revision number). This can help establish blame in the event a change caused a malfunction, or identify the author of brilliant code.

A special case is the Java programming language, where annotations can be used as a special form of syntactic metadata in the source code. Classes, methods, variables, parameters and packages may be annotated. The annotations can be embedded in class files generated by the compiler and may be retained by the Java virtual machine and thus influence the run-time behavior of an application. It is possible to create meta-annotations out of the existing ones in Java, which makes this concept more sophisticated than in other languages like C#.

In the digital imaging community the term annotation is commonly used for visible metadata superimposed on an image without changing the underlying master image, such as sticky notes, virtual laser pointers, circles, arrows, and blackouts (cf. redaction).

In the medical imaging community, an annotation is often referred to as a region of interest and is encoded in DICOM format.

In the United States, legal publishers such as Thomson West and Lexis Nexis publish annotated versions of statutes, providing information about court cases that have interpreted the statutes. Both the federal United States Code and state statutes are subject to interpretation by the courts, and the annotated statutes are valuable tools in legal research.

In linguistics, annotations include comments and metadata; these non-transcriptional annotations are also non-linguistic. A collection of texts with linguistic annotations is known as a corpus (plural corpora). The Linguistic Annotation Wiki describes tools and formats for creating and managing linguistic annotations.

A web annotation is an online annotation associated with a web resource, typically a web page. With a Web annotation system, a user can add, modify or remove information from a Web resource without modifying the resource itself. The annotations can be thought of as a layer on top of the existing resource, and this annotation layer is usually visible to other users who share the same annotation system. In such cases, the web annotation tool is a type of social software tool. For Web-based text annotation systems, see Text annotation.

Web annotation can be used for the following purposes:

- To rate a Web resource, such as by its usefulness, user-friendliness, suitability for viewing by minors.
- To improve or adapt its contents by adding/removing material, something like a wiki.
- As a collaborative tool, e.g. to discuss the contents of a certain resource.
- As a medium of artistic or social criticism, by allowing Web users to reinterpret, enrich or protest against institution or ideas that appear on the Web.
- To quantify transient relationships between information fragments.

2. EXISTING SYSTEM

An increasing number of databases have become web accessible through HTML form-based search interfaces. The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the en-

coded data units to be machine processable, which is essential for many applications such as deep web data collection and Internet comparison shopping, they need to be extracted out and assigned meaningful labels. In this paper, we present an automatic annotation approach that first aligns the data units on a result page into different groups such that the data in the same group have the same semantic. Then, for each group we annotate it from different aspects and aggregate the different annotations to predict a final annotation label for it. An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database. Our experiments indicate that the proposed approach is highly effective. [1]

In this paper, we studied the data annotation problem and proposed a multiannotator approach to automatically constructing an annotation wrapper for annotating the search result records retrieved from any given web database. This approach consists of six basic annotators and a probabilistic method to combine the basic annotators. Each of these annotators exploits one type of features for annotation and our experimental results show that each of the annotators is useful and they together are capable of generating highquality annotation. A special feature of our method is that, when annotating the results retrieved from a web database, it utilizes both the LIS of the web database and the IIS of multiple web databases in the same domain. We also explained how the use of the IIS can help alleviate the local interface schema inadequacy problem and the inconsistent label problem.[1]

In our paper, we also studied the automatic data alignment problem. Accurate alignment is critical to achieving holistic and accurate annotation. Our method is a clustering based shifting method utilizing richer yet automatically obtainable features. This method is capable of handling a variety of relationships between HTML text nodes and data units, including one-to-one, one-to-many, many-to-one, and one-to-nothing. Our experimental results show that the precision and recall of this method are both above 98 percent. [2] There is still room for improvement in several areas. For example, we need to enhance our method to split composite text node when there are no explicit separators. We would also like to try using different machine learning techniques and using more sample pages from each training site to obtain the feature weights so that we can identify the best technique to the data alignment problem. [2]

An increasing number of databases have become Web accessible through HTML form-based search interfaces. The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the encoded data units to be machine processable, which is essential for many applications such as deep Web data collection and comparison shopping, they need to be extracted out and assigned meaningful labels. In this paper, we present a multiannotator approach that first aligns the data units into different groups such that the data in the same group have the same semantics. Then for each group, we annotate it from different aspects and aggregate the different annotations to predict a final annotation label. An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same site. Our experiments indicate that the proposed approach is highly effective. [3]

In this paper, we studied the data annotation problem and proposed a multi-annotator approach to automatically constructing an annotation wrapper for annotating the search result records retrieved from any given Web database. Each of these annotators exploits one type of special features for annotation and our experimental results indicate that each proposed annotator is useful and they together are capable of generating high quality annotation wrappers. We also illustrated how the use of the integrated interface schema can help alleviate the local interface schema inadequacy problem and the inconsistent label problem. In addition, a new data alignment technique using richer yet automatically obtainable features was proposed to cluster data units into different groups/concepts in support of more robust and holistic annotation. There is still room for improvement in several areas as mentioned in Section 6. For example, we need to enhance the ability to deal with multi-valued attributes that may have more than one value for some SRR (e.g., authors for books). [3]

There are many available methods to integrate information source reliability in an uncertainty representation, but there are only a few works focusing on the problem of evaluating this reliability. However, data reliability and confidence are essential components of a data warehousing system, as they influence subsequent retrieval and analysis. In this paper, we propose a generic method to assess data reliability from a set of criteria using the theory of belief functions. Customizable criteria and insightful decisions are provided. The chosen illustrative example comes from real-world data issued from the Sym'Previus predictive microbiology oriented data warehouse.

We proposed a generic method to evaluate the reliability of data automatically retrieved from the web or from electronic documents. Even if the method is generic, we were more specifically interested in scientific experimental data. [4]

The method evaluates data reliability from a set of common sense (and general) criteria. It relies on the use of basic probabilistic assignments and of induced belief functions, since they offer a good compromise between flexibility and computational tractability. To handle conflicting information while keeping a maximal amount of it, the information merging follows a maximal coherent subset approach.

Finally, reliability evaluations and ordering of data tables are achieved by using lower/upper expectations, allowing us to reflect uncertainty in the evaluation. The results displayed to end users are an ordered list of tables, from the most to the least reliable ones, together with an interval-valued evaluation. [4]

We have demonstrated the applicability of the method by its integration in the @Web system, and its use on the Sym'Previus data warehouse. As future works, we see two main possible evolutions:

Complementing the current method with useful additional features: the possibility to cope with multiple experts, with criteria of no equal importance and with uncertainly known criteria;

Combining the current approach with other notions or sources of information: relevance, in particular, appears to be equally important to characterize experimental data. Also, we may consider adding user feedback as an additional (and parallel) source of information about reliability or relevance, as it is done in web applications. [4]

With XML becoming a ubiquitous language for data interoperability purposes in various domains, efficiently querying XML data is a critical issue. This has lead to the design of algebraic frameworks based on tree-shaped patterns akin to the tree-structured data model of XML. Tree patterns are graphic representations of queries over data trees. They are actually matched against an input data tree to answer a query. Since the turn of the 21st century, an astounding research effort has been focusing on tree pattern models and matching optimization (a primordial issue). This paper is a comprehensive survey of these topics, in which we outline and compare the various features

Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

of tree patterns. We also review and discuss the two main families of approaches for optimizing tree pattern matching, namely pattern tree minimization and holistic matching. We finally present actual tree pattern-based developments, to provide a global overview of this significant research topic. [4]

We provide in this paper a comprehensive survey about XML tree patterns, which are nowadays considered crucial in XML querying and its optimization. We first compare TPs from a structural point of view, concluding that the richer a TP is with matching possibilities, the larger the subset of XQuery/XPath it encompasses, and thus the closer to user expectations it is.

Second, acknowledging that TP querying, i.e., matching a TP against a data tree, is central in TP usage, we review methods for TP matching optimization. They belong to two main families: TP minimization and holistic matching. We trust we provide a good overview of these approaches' evolution, and highlight the best algorithms in each family as of today. Moreover, we want to emphasize that TP minimization and holistic matching are complementary and should both be used to wholly optimize TP matching. [4]

3. PROPOSED ALGORITHM

Annotating the web search results are very useful these days for the users for many reasons such as comparison of available products, online shopping, articles etc.

This work is proposing a mechanism for fast searching of web data for articles using annotations applied automatically for future usage for the various websites visited by the users of the system. The complete algorithm shall be implemented using following steps:

Step 1:
Load the website in the system
Step 2:
Retrieve the articles from it using various links available on the pages of the website
Step 3:
Look for heading tags, bold/strong tags, and frequency of the words in the articles to decide the annotation for the specific articles
Step 4:
Maintain the list of the various articles and provide them as quick link lists for future usage
Step 5:

Update the annotation list using every new website visited using the above steps

1. Algorithms Applied for annotating the articles:
- Data Unit and Text Node Feature Extraction
- Data Content
- Presentation Style
- Data Type
- Tag Path
- Adjacency

2 Data Alignment Techniques
3 Annotators
- Frequency Based Annotators
- Schema Value Annotators
- Common Knowledge Annotators

4. FUTURE WORK

The proposed work is being implemented using C# for websites of the blogs where annotation is required to a great extent using standard machines.

REFERENCES

[1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Clement Yu, "Annotating Search Results from Web Databases", Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 3, March 2013

[2] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Clement Yu, "Annotating Structured Data of the Deep Web", This work is supported in part by the following NSF grants: IIS-0414981, IIS-0414939 and CNS-0454298.

[3] Sebastien Destercke, Patrice Buche, and Brigitte Charnomordic, "Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse", Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 1, January 2013

[4] Marouane Hachicha and Je´ roˆme Darmont, "A Survey of XML Tree Patterns", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013

[5] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. SIGMOD Int'l Conf. Management of Data, 2003.

[6] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large Web Sites," Proc. Sixth Int'l Workshop the Web and Databases (WebDB), 2003.

[7] P. Chan and S. Stolfo, "Experiments on Multistrategy Learning by Meta-Learning," Proc. Second Int'l Conf. Information and Knowledge Management (CIKM), 1993.

[8] W. Bruce Croft, "Combining Approaches for Information Retrieval," Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, Kluwer Academic, 2000.

[9] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites," Proc. Very Large Data Bases (VLDB) Conf., 2001.

[10] S. Dill et al., "SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation," Proc. 12th Int'l Conf. World Wide Web (WWW) Conf., 2003.

[11] H. Elmeleegy, J. Madhavan, and A. Halevy, "Harvesting Relational Tables from Lists on the Web," Proc. Very Large Databases (VLDB) Conf., 2009.

[12] D. Embley, D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y. Ng, and R. Smith, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31, no. 3, pp. 227-251, 1999.

[13] D. Freitag, "Multistrategy Learning for Information Extraction," Proc. 15th Int'l Conf. Machine Learning (ICML), 1998.

[14] D. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, 1989.

[15] A. Arasu and H. Garcia-Molina. Extracting Structured Data from Web pages. SIGMOD Conference, 2003.

[16] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo. Automatic Annotation of Data Extracted from Large Web Sites. WebDB Workshop, 2003.

[17] P. Chan and S. Stolfo. Experiments on Multistrategy Learning by Meta-Learning. CIKM Conference, 1993.

[18] W. Bruce Croft. Combining approaches for information retrieval. In Advances in Inf. Retr.: Recent Research from the Center for Intel. Inf. Retr., Kluwer Academic, 2000.

[19] V. Crescenzi, G. Mecca, and P. Merialdo. RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites. VLDB Conference, 2001.